Available online at www.sciencedirect.com
ScienceDirect
Procedia Computer Science 35 (2014) 894 – 901
Procedia Computer Science

18th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems - KES2014

Evaluation of communication and travel behavior extraction with latent topics

Nobuo Suzuki[a*], Kazuhiko Tsuda[b]

[a]KDDI Corporation, Iidabashi 3-10-10, Chiyoda, Tokyo 102-8460, Japan
[b]University of Tsukuba, Otsuka 3-29-1, Bunkyo, Tokyo 112-0012, Japan

Abstract

This study proposed the habitual behavior information extraction method from the data on Internet to build effective behavioral change support system so far. It is well known that habitual behavior improvement is important to avoid risk behaviors for a safety driving and a health improvement. It used Latent Dirichlet Allocation approach and evaluated by using telecommunication behaviors in Question and answering Web sites. This paper describes another evaluation by using travel behavior information. On the other hand, the dependency relation is often used to extract valuable information from text data. It also shows the comparative evaluation between our proposed method and the dependency relation method. It is realized the proposed method is more accurate than the dependency relation method according to the result of the evaluation.

Peer-review under responsibility of KES International.

Keywords: Habitual behavior; Behavioral modification; LDA; PMI

1. Introduction

Large text data are generated daily at SNS and Question and answering sites on the Internet [14]. The method for extraction habitual behavior information from such data has been proposed to build a behavioral modification supporting system affected to safety driving and health improvement [9]. It is well known that the improvement of habitual behavior is significantly important for a habitual behavior subject to improve health by prohibiting smoking and avoiding a dangerous behavior such as using a mobile phone while driving. Therefore, this study has tried to extract the information focused on habitual behaviors.

* Corresponding author
E-mail address: suzu3nobu@gmail.com

Peer-review under responsibility of KES International.
doi:10.1016/j.procs.2014.08.163

Actual extraction method of habitual behaviors presumed latent topics included in text by Latent Dirichlet Allocation, LDA, which was one of topic models and decided the words suited for habitual behaviors by Point-wise Mutual Information, PMI, from candidate words included the topics. This method is called the proposed method in this paper. This paper reports the result of the evaluation experiment at travel behaviors by using transport facilities on Question and answering sites, in addition to the evaluation experiment at telecommunication behaviors so far. The dependency relation is frequently used as a method to extract the valuable information from text data. Therefore, our proposed method has been evaluated by comparing with the dependency relation method with data same as the evaluation experiment of the proposed method before. It was realized that the proposed method obtained higher accuracy rate than the dependency relation method by the result of the comparative evaluation.

2. Methods of extracting behavior information with latent topics and PMI

2.1. How the extraction method works

The proposed method extracts the habitual behaviors information with LDA and PMI as below. First, the habitual behavior was defined as frequently appearing human's common behavior that was not only physiological habitats such as tooth brushing and sleeping. The habitual behavior, therefore, includes three elements those are an action, an object and periodic frequency information. The habitual behavior HB is defined as combination of the formula (1).

$$HB = \{Frequency, Action, Object\} \tag{1}$$

The habitual behaviors were extracted by using LDA that was one of the topic model technologies [3]. The feature of the topic model is to express one document as a mixture of one or more topics. Canini et. al showed it could model documents with high accuracy. Our method prepared frequently keywords used as periodical expressions such as "Yoku", "Mai" and "Itsumo". A morpheme analysis is carried out by using the extracted sentences. It selects bag-of-words with adjectives, verbs, nouns and adverbs that are easily used as objects of the frequency, the action and the object. Then, LDA processing is executed. As a result, the topics constructed with one or more words are extracted and then the topics that have periodic expressions in those topics are also extracted. Some words are included at the time. They become the candidates of actions and objects expressed the habitual behaviors other than the periodic expressions in extracted each topics. The habitual behaviors, however, cannot be extracted more accurately because extracted topics include some words other than habitual behaviors in extracted topics even as they are. Therefore, the method assumes the words related to habitual behaviors with PMI as an index from the words in the topics. Then, the keywords are applied to "Frequency" of habitual behaviors. PMI between the words for "Action" and frequency keyword is calculated. They include Verb-independent, Noun-Sa-changing and Noun-adverb-available. Then two largest words are extracted as Action. It calculates PMI between the nouns and the periodical keywords, and then selects the best three words. The nouns includes Noun-Sa-changing that wasn't selected at Action and excludes Noun-independent. PMI is calculated with formula (2) and (3). It is an index that can express the strength among words and shows strength of relation between Action and Object words with the periodical words.

$$PMI(x,y) = \log_2 \frac{p(x,y)}{p(x)p(y)} \tag{2}$$

$$p(x) = \frac{f(x)}{N},\ p(y) = \frac{f(y)}{N},\ p(x,y) = \frac{f(x,y)}{N} \tag{3}$$

2.2. Evaluation by using travel behaviors

The former study used 8,953 text sentences of Question and answering sites and evaluated this method in telecommunication behaviors. 20 topics were extracted and 18 topics were decided to right habitual behaviors as a result. The accuracy was 90%.

The travel behaviors other than telecommunication behaviors also were evaluated at this time. The behavioral change promoted to use public transportation systems other than individual automobiles as a countermeasure against traffic jam and global warming is needed and becomes an object for behavioral change. The questions and answers regarding transportations include a lot of habitual behaviors related to movement of persons. It is possible to extract habitual behaviors from question answering text data about transportations and utilize to behavioral change with the information. 6,627 sentences of utterances are collected from question answering sites regarding transportations. The sentences included keywords of periodical expressions same as the telecommunication behaviors are extracted from them. 521 sentences are extracted by this process and analyzed by the morpheme analysis tool [2]. Next, each word of adjectives, verbs, nouns and adverbs are extracted. Then, topics are extracted every utterance by LDA tool, LDA-C. Table 1 shows the examples of the travel behavior topics in the question answering sites. 16 topics were obtained by selecting keywords of periodical expressions from the topics.

Table 1. The Examples of Topics in Question Answering Sites (Travel Behaviors)

Topic	Words (Parts of speech)
Topic 000	Itsumo (Adverb), Ari (Verb-Dependence), Basu (Noun), I (Verb-Dependence), Shi (Verb-Independence), En (Noun-Suffix-Numeral), Ryokin (Noun), Hi (Noun-Suffix-Numeral), Riyou (Noun-Sa-Change-Connection), Nagahara (Proper noun-Person's name-Last name)
Topic 026	Yoku (Adverb), No (Noun-Dependence) Basu (Noun), Adobencha-wa-rudo (Noun), Tore (Verb-Dependence), Hi (Noun-Suffix-Numeral), Me (Noun-Suffix), Shi (Verb-Dependence), Jikan (Noun-Adverb available), Asa (Noun-Adverb available)

The candidates were selected by PMI values at each topic and the words suitable to habitual behaviors were extracted. Table 2 shows the examples of the habitual behaviors selected from topics in the case of travel behaviors.

Table 2. The Examples of the Habitual Behaviors Selected from Topics in the Case of Travel Behaviors.

Topic	Frequency	Action	Object	Explanations
Topic 000	Itsumo (Always)	Ari (Exist)	Basu (Bus)	I always use a bus.
		Shi (Do)	Riyou (Usage)	
Topic 026	Yoku (Often)	Shi (Do)	Basu (Bus)	I often go to Adventure World at the morning by a bus.
		Jikan (Time)	Adobencha-wa-rudo (Adventure World)	
			Asa (Morning)	

Then, all topics were confirmed that the word extracted by the proposed method expressed habitual behaviors manually. It was able to decide 12 topics out of 16 extracted properly and the accuracy was 75%.

This evaluation experiment couldn't obtain the habitual behaviors from the words extracted at four topics. Table 3 shows these examples. All incorrect topics couldn't extract right words as Action and Object. They also included the words unrelated to transportation such as last names and "Meigi (ownership)", and news unrelated to transportation after inspection the data. Using the words same as place names in most Japanese last names causes that the topics include those words. "Kobuchizawa", for instance, has different meanings as both a land name and a last name in Japan. Therefore, a last name would be applied by relating to the word "Meigi", and a place name would be applied by relating to travelling. It is necessary to adapt the detection mechanism between place names and last names according to their contexts. The countermeasure is available such as deleting the words that have a low relationship with transportation by word frequency information at pretreatment of data. The information of time and place is important for travelling. It is necessary to select preferentially the words related to time and place when they are included in the extracted words.

Table 3. The Examples of Errors (Travel Behaviors).

Topic	Frequency	Action	Object	The reasons of errors
Topic 011	Yoku (Often)	Shi (Do)	Kyoto	Extracting actions is incorrect. There aren't any correct words in other words in the topics.
			Aoshima	
			Hokkaido	
Topic 035	Yoku (Often)	Waribiki (Discount)	---	Extracting objects is incorrect. There aren't any correct words in other words in the topics.
		Iku (Go)		
Topic 049	Yoku (Often)	Ari (Exist)	Kobuchizawa	Extracting objects is incorrect. Travelling and "Meigi" is not related.
		Shi (Do)	Meigi	

3. Extraction of the habitual behavioral information by dependency relation

3.1. Extraction method

The dependency relation is one of the famous methods to extract valuable information from text data. Itoh et. al extracted chronological patterns of actions and interests from blogs [7]. Endo et. al also extracted emotional expressions [4]. Furthermore, Ikeda et. al proposed the harmful information detecting method by dependency relations so far [5]. The habitual behavior extraction method with dependency relations has been realized and evaluated by comparison with proposed method.

Itoh's system analyzed the time series transition of the action descriptions for the objects each month such as "to be infected with new influenza virus". It also showed the tree structures of the relation between the objects and the actions. Then, it visualized those time series transition with 3-dimensional space. They implemented the topic searching system that visualized and searched the time series transition of actions, interests and impressions for the objects written on the blogs. This system extracted the dependency relation between nouns and verbs as events of the object and the action from the result of the dependency relation analysis. It classified the semantic relations such as agent's actions and objects by the case-marking particles belonging to nouns, and made into a database. Next, Endoh's method extracted the emotional generating expressions as the steps below. To make the emotional expressions dictionary for seeds, collect the candidates of the generating expressions, and filter the candidates of the generating expressions. It took out the sentences including the expressions registered in the emotional expressions dictionary for seeds from the corpus, and parsed them. It then extracted two modifier phrases of the emotional expressions from the results, and decided the expressions had the ends of them including only "noga" and "kotoga" to the candidates of the emotional generating expressions. Three modifier phrases of the emotional expressions become the candidates of the emotional generating expressions if "kotoga" shows the end of expressions. It filtered them after that. The candidates of the emotional generating expressions aren't become if the top morpheme of the first phrase is adnominal adjective, nouns-independent or noun-number. It extracted the emotional generating expressions which the top morpheme of their second phrase was adjectives-dependent, verbs, nouns-Sa-changing-connection, or nouns-adjective-verb-stem. At last, Ikeda proposed the machine learning method with illegality and harmfulness by extracting the pair of phrases dependency related from documents. It also detected illegal and harmful information with high accuracy by abstracting and expanding the pair of the phrases using the conceptual dictionary. The error detection of the ordinary method was corrected and the accuracy was improved by detecting the pair of the sentences dependency related with high illegality and harmful from the sentences with deciding to illegality, harmful and harmless at the ordinary method. This method also extracted many expressions by abstracting and expanding the pair of dependency related phrases by using the conceptual dictionary.

Figure 1 shows the extraction method with dependency relations realized at this time. First, the sentences included periodical expressions same as mentioned in Chapter 2 from text data are selected. The keywords of these periodical expressions are defined as "Frequency". Next, the syntax analysis processing is performed and the dependency relations of each phrase are obtained. The verbs appeared in the phrase depended to the periodical expressions by obtained dependency relations are applied to "Action". The nouns in the phrase depended to that depended phrase among the phrase appeared in the first periodical expression and the depended phrase are applied to "Object" of the habitual behavior. The first phrase is the extracted object when there are one or more phrases for actions and objects.

3.2. Evaluation experiment

The communication and travel behaviors were evaluated by using data same as evaluation of proposed method in the evaluation of dependency relations. CaboCha was used as the dependency relation analysis tool [8].

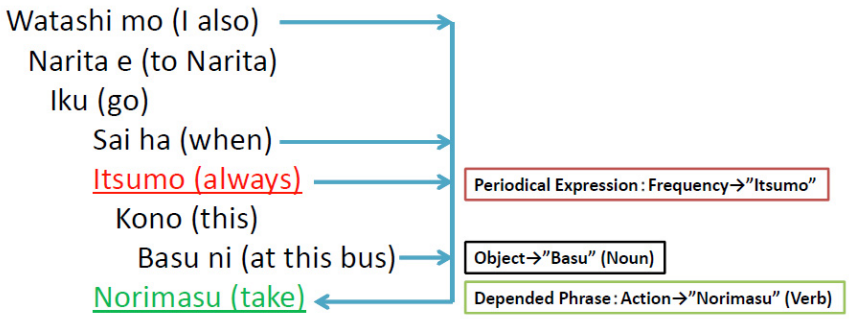

Figure 1. The Habitual Behavior Extracting Procedure with Dependency Relation (The parentheses show English translation).

First, the dependency relation method for telecommunication behaviors was evaluated by using text data 8,953 sentences on the question answering sites of the telecommunication company. It could recognize 153 habitual behaviors and 63 sentences were extracted correctly from them as a result. The accuracy was 41.2% according the experiment. Table 4 shows the correct examples and table5 shows the incorrect examples.

Table 4. The correct examples of the dependency relation method. (Telecommunication behaviors)

No.	Frequency	Action	Object	Explanations
1	Yoku (Often)	Kakeru (Call)	Aite (The person at the other end of the line.)	I often call the person.
2	Itsumo (Always)	Tsukau (Use)	Teigaku (Required amount money)	I always use the required amount.
3	Yoku (Often)	Machigaeru (Mistake)	BCC	I often mistake BCC.

Table 5. The incorrect examples of the dependency relation method. (Telecommunication behaviors)

No.	Frequency	Action	Object	Reasons of incorrect
1	Yoku (Often)	Suru (Do)	Gaso (Pixels)	Extraction of the object is incorrect.
2	Mai (Every)	Suru (Do)	Samazama (Various)	Extraction of the object is incorrect.

3	Itsumo (Always)	Tasukaru (Helpful)	*	Extraction of the object is incorrect.

521 sentences included periodical expressions same as the keywords of telecommunication behaviors were extracted by using 6,627 sentences on the question answering sites related to transportation same as the evaluation of proposed method. The evaluation of the dependency relation with travel behaviors extracted 86 habitual behaviors. 29 behaviors were recognized to extract correctly in those behaviors. Then, the accuracy was 33.7%. Table 6 and 7 show the correct and incorrect examples respectively.

Table 6. The correct examples by using the dependency relation. (Travel behaviors)

No.	Frequency	Action	Object	Explanations
1	Itsumo (Always)	Noru (Take)	Basu (Bus)	I always take a bus.
2	Yoku (Often)	Tomeru (Park)	Kinrin (Neighborhood)	I often park at neighborhood.
3	Yoku (Often)	Shiraberu (Check)	Densha (Train)	I often check trains.

Table 7. The incorrect examples by using the dependency relation. (Travel behaviors)

No.	Frequency	Action	Object	Explanations
1	Itumo (Always)	Suru (Do)	Hyaku (Hundred)	Extraction of the object is incorrect.
2	Yoku (Often)	Wakaru (Understand)	Kata (For)	Extraction of the object is incorrect.
3	Yoku (Often)	Iku (Go)	Kata (Shoulder)	Extraction of the object is incorrect.

The reason of incorrect result is to extract the words have no meaning as “Object” and it is necessary that the word dictionaries and the extraction of proper nouns are simultaneously used. On the other hand, regarding the accuracy between the dependency relation and the proposed method, the latter has more accurate than the former at the telecommunication behaviors and the travel behaviors both. Therefore, it is able to understand that the proposed method is more effective way than the dependency relation. Unfortunately, there are issues about both methods in common like lacking of information for behavioral modifications. The information of when, where and who are insufficient in case of extraction the habitual behavior like “I always take a bus.” at both methods. The conditions of behaviors also are needed to connect to the behavioral modification. Expansion of the habitual behavior definition and information extraction by frames is able to handle those issues.

4. Conclusion

This paper reported the evaluation of the already proposed method that used LDA and PMI. The telecommunication behavior was evaluated only so far. This study added the evaluation of the travel behavior data and realized high accuracy. Furthermore, the proposed method was evaluated by comparing with the dependency relation that frequently used in information extraction from text data. It was realized that the proposed method had higher accuracy in both of the telecommunication and travel behavior evaluations as

showed Table 8.

The future tasks are that more habitual behaviors will be collected by expanding to the health improving behaviors and the content of extracting information also will be enriched.

Table 8. The comparison result of LDA&PMI method and the dependency relation.

.	LDA&PMI	Dependency Relation
Telecommunication Behaviors	90.0%	41.2%
Travel Behaviors	75.0%	33.7%

References

[1] Bianchi A., Phillips GJ. Psychological Predictors of Problem Mobile Phone Use, *Cyberpsychology & Behavior*; 2005, Vol.8, No.1, pp.39-51

[2] Blei MD., Ng YA., Jordan IM. Latent Dirichlet Allocation, *Journal of Machine Learning Research*; 2003, Vol. 3, pp.993-1022

[3] Canini RK., Shi L., Griths LT. Online Inference of Topics with Latent Dirichlet Allocation, *Proceeding of the 12th International Conference on Artificial Intelligence and Statistics*; 2009

[4] Endoh D., Saitoh M., Yamamoto K. Extraction the expressions of occurring emotions with dependency relations, *NLP2006*; 2006, pp.947-950.

[5] Ikeda K., Yanagihara T., Matsumoto K., Takishima Y. Detection of Illegal and Hazardous Information Using Dependency Relations and Keyword Abstraction, *DEIM Forum*; 2010, C9-5

[6] Inui K., Abe S., Morita H., Eguchi M., Sumida A., Sao C., Hara K., Murakami K., Matsuyoshi S. Experience Mining: Building a Large-Scale Database of Personal Experiences and Opinions from Web Documents. *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*; 2008, pp.314-321

[7] Itoh M., Yoshinaga N., Toyoda M., Kitsuregawa M. 3D Visualization System for Time Sequential Changes in Blog Users' Activities and Interests, *DEIM Forum*; 2011, E7-5

[8] Kudo T. and Yuji Matsumoto: Japanese Dependency Analysis using Cascaded Chunking, CONLL 2002

[9] Kukkonen OH. Behavior Change Support Systems: The Next Frontier for Web Science, *Proceedings of the Second International Web Science Conference*; 2010

[10] Fujikura T., Fujimura K., Okuda H. Mining Experiences from Large-Scale Texts, *IEICE Transactions on Information and Systems*; 2009, Vol.J92-D, No.3, pp.301-310

[11] Takahashi M., Satoh S., Matsuo M. Human Behavior Process Extraction from the Web, *IEICE SIG-AI*; 2012, Vol.112, No.319, AI2012-20, pp.31-35

[12] Tanaka S., Nakamura K., Teraguchi T., Nakamoto S., Kato R. Research on Analysis of Users' Habitual Behavior in Microblog, *The 75th National Convention of IPSJ*; 2013, No.5N-4

[13] Suzuki N., Fujita Y., Tsuda K., Effective Extraction Method of Loss Aversion Utterances based on the Expected Utility, *Frontiers in Artificial Intelligence and Applications, Advances in Knowledge-Based and Intelligent Information and Engineering Systems*; 2012, IOS Press, Vol.243, pp.833-840

[14] Suzuki N. and Tsuda K.: An Effective Method for Habitual Behavior Extraction from the Internet, Procedia Computer Science, Vol.22, pp.599-605, 2013

[15] Bianchi A., Phillips GJ., Psychological Predictors of Problem Mobile Phone Use, *Cyberpsychology & Behavior*; 2005, Mary Ann Liebert, Inc, Vol.8, No.1

[16] Redelmeier AD., Tibshirani JR., Association between cellular-telephone cells and motor vehicle collisions, *The New England Journal of Medicine*; 1997, Vol.336, No.7, pp.453-458